International Journal of Applied Information Systems (IJAIS) – ISSN : 2249-0868
Foundation of Computer Science FCS, New York, USA
Volume 5– No.1, January 2013 – www.ijais.org

N-layer Approach to Web Information Retrieval

Jayant Gadge
Research Scholar, VJTI
Matunga, Mumbai
India

S.S. Sane, Ph.D
Professor, VJTI
Matunga, Mumbai
India

H.B. Kekre, Ph.D
Sr. Professor, MPSTME
SKVM's NMIMS University
Mumbai, India

ABSTRACT

In web information retrieval, the terms or keywords are used for indexing purpose of document. These terms or keywords appear in special location such as title, subtitle, header, hyperlinks and so on. Vector space model ignores the importance of these terms with respect to their position while calculating the weight of the indexing terms. The effectiveness of the vector space model crucially depends on the weights applied to the terms of the document vectors. These weights are found using a term weight evaluation scheme based on the frequency of the terms in the document and the collection. Terms that occur more often in a document are treated as more important whereas terms that occur less frequently throughout a collection are given a higher weight.

In N-level Vector space approach, the importance of these terms with respect to their position is considered. The web document is logically divided in N-layer considering the structure of web document and weights are assigned to terms based on their presence in different layer within the document. Different weight evaluation schemes proposed for vector space models are applied to N-level vector space model and are compared. N-layer vector space model gives better result as compare to vector space model. Cosine similarity and all six weight evaluation methods that are formed using different local weights and global weights show that average precision and average recall in case of N-layer vector space model is always better than vector space model.

General Terms

Web Information Retrieval, Web Mining

Keywords

N-layer vector space model, global weight, local weight, weight evaluation scheme.

1. INTRODUCTION

The goal of an Information Retrieval is to help a user to locate the most similar documents that have the potential to satisfy the user information needs. To solve this problem, researchers have proposed several models. The focus of information retrieval is to search for information relevant to a user's needs within a collection of data which is relevant to the user's query [1] [2] [3]. User will formulate query and send the query to the search engine. Search engine searches for the matches in the document dataset and retrieves results. The user will evaluate the results based on the relevance. If the user feels that it is a relevant document, he finishes the search else user continues to search in the dataset by reformulating the query until the relevant documents are retrieved [4] [5].

A very popular IR model [6] used in recent years has been the vector space model. It has been observed that the effectiveness of the vector space model depends on the term weights applied to the terms of the document vectors. These term weights are found using a term weight evaluation scheme based on the frequency of the term in the document as well as frequency of term in the collection. Terms that occur more often in a document are treated as more important, i.e. they better describe the document content and are given a higher weight. Terms that occur less frequently throughout a collection are given a higher weight because it assumed that these terms better differentiate between documents. Measures derived from knowledge of the document are regarded as local measures, and measures derived from knowledge of the collection are regarded as global measures.

The vector space representation has limitation. First, the order in which the terms appear in the document is lost in the vector space representation and second, terms are statistically independent. To overcome these limitations, N-layer vector space model is proposed. In N-layer vector space representation, semi-structured characteristics of web document are considered. The terms that appearing in the special locations such as title, hyperlinks, body and paragraph represent more important information in the web document. The document is logically divided in N-layers according to the structure and weights are assigned to terms based on their presence in different layer within the document.

In this paper, all related work in web information retrieval is discussed in section 2. In Section 3, N-level vector space model and algorithm is described. In Section 4, Weight evaluation scheme which is used for comparison of vector space model and n-layer vector space model is discussed. In Section 5, experimental results and graphs are discussed and in Section 6, the conclusion is presented

2. RELATED WORK

The task of an information retrieval system is to identify relevant documents based on a user's information need. Over the past decades, many different retrieval models have been proposed, studied and empirically validated. In order to make information retrieval efficient, the documents are typically transformed into a suitable representation. There are several representations such as Boolean Retrieval model, Fuzzy Set model, Extended Boolean model, Vector Space model, Latent Semantic indexing model. A brief survey of these methods is been given below.

Boolean model [7] is based on set theory and Boolean algebra. The document is represented as set of terms and queries are as Boolean expressions formed using terms. Boolean model retrieves the document if there is exact match between query and set of document. Sometimes it leads to either too many or too few retrieved documents. This model is easy to implement. This model has limitation for ranking documents and assigning importance factors or weights to query terms.

The Latent Semantic Indexing [8] information retrieval model builds upon the prior research in information retrieval and using the singular value decomposition [9]. It reduces the dimensions of the term-document space and attempts to solve the synonymy and polysemy problems that affect automatic information retrieval systems. LSI explicitly represents terms and documents in a rich, high-dimensional space, allowing the underlying semantic relationships between terms and documents to be exploited during searching. The process of matching documents could be based on concept matching instead of index term matching. The main idea in this model is to map each document and query vector into a lower dimensional space which is associated with concepts.

The probabilistic retrieval model [10] is based on the probability ranking principle which states that an information retrieval system is supposed to rank the documents based on their probability of relevance to the query, given all the evidence available. The principle takes into account that there is uncertainty in the representation of the information need and the documents. There can be a variety of sources of evidence that are used by the probabilistic retrieval methods. The most common one is the statistical distribution of the terms in both the relevant and non-relevant documents. In 1976, Stephen Robertson and Karen Sparck Jones [11] proposed a probabilistic model for information retrieval based on Independence Assumptions and ordering Principles for probable relevance.

3. RELATED WORK

In N-layer vector space representation, semi-structured characteristics of web document are considered. The terms that appearing in the special locations such as title, hyperlinks, body and paragraph represent more important information in the web document. The document is logically divided in N-layers according to the structure and weights are assigned to terms based on their presence in different layer within the document. The document is logical divided in three layers namely Title region, hyperlink region and body region. The algorithm used for N-layer vector space model is given below.

```
Step1:
Let D = {$D_1$, $D_2$, $D_3$… $D_N$ } be the web document set

 Represent the document in vector space
   For each $D_k$ in the Document set
      Remove the Stop word
      Find all tokens in title region
      Find all token in the body region
      Find all the Hyperlinks and Anchors in the $D_k$
      Apply Porter's Stemming algorithm for remove
      morphological variants of all terms
   End for

Step 2:
     Prepare Dictionary of terms

Step 3:
     The documents are represented in N- layer vector space
     using sparse matrix and document is divided in the
     following regions.
               Title Region
               Hyperlinks region
               Body region
Step 4:
     Calculate term frequency  $tf_{ik}$ for term using
```

$$tf_{ik} = \propto \times\ tf_{ik1} \times \log\left(\frac{M}{tf_{ik1}}\right) + \beta\ \times\ tf_{ik2} \times log\left(\frac{M}{tf_{ik2}}\right) + \gamma\ \times\ tf_{ik3} \times log\left(\frac{M}{tf_{ik3}}\right)$$

α – Weight assigned to term appearing in Title region
β - Weight assigned to term appearing in Hyperlink region
γ – Weight assigned to term appearing in body Region
α > β > γ here α= 2, β = 1.5 and γ = 1

M – Sum of term frequency of all region
tf_{ik1} – terms that appearing in title region
tf_{ik2} - terms that appearing in hyperlink region
tf_{ik3} - terms that appearing in body region
$M = tf_{ik1} + tf_{ik2} + tf_{ik3}$

Step 5:

Calculate Inverse document frequency idf for term using

$$idf_i = log\frac{N}{n_i}$$

Where
idf_i – denote inverse document frequency of term i
N – Total number of document in the collection
n_k -- The number of document in the collection that contain term t_i

Step 6:

Calculate weight of term w = tf * idf

$$w_i \frac{tk_{ik}\ \log(N/n_i)}{\sqrt{\sum_{i=1}^{t}(tf_{ik})^2\ [(log\,N/n_i)]^2}}$$

Step 7:

Calculate similarity between document and query sim (Di, q) using cosine similarity

$$sim(D_j, q) = \frac{\sum_{j=1}^{t} w_{qj}\ *\ w_{ij}}{\sqrt{\sum_{i=1}^{t}(w_{iq})^2\ \sum_{i=1}^{t}\left(w_{d_{ij}}\right)^2}}$$

4. TERM WEIGHT

Proper term weight evaluation can greatly improve the performance of the vector space method. A weight evaluation scheme is composed of three different types of term weight evaluation component: local weight, global weight, and normalization [13][14][15]. The term weight is given by

$L_{ik}\, G_i\, N_j$

Where L_{ik} is the local weight i.e. term frequency tf for term i in document k
G_i is the global weight i.e. inverse document frequency for term i and
N_k is the normalization factor for document k.

Local weights are functions of how many times each term appears in a document, global weights are functions of how many times each term appears in the entire collection, and the normalization factor compensates for discrepancies in the lengths of the documents.

4.1 Local Weight

The local weight [13][14][15] in the given document is simply the number of times a given term appears in that document. This count is normalized to prevent a bias towards longer documents which may have a higher term count regardless of the actual importance of that term in the document. To measure importance of the term t within the particular document D, three approaches are considered that is shown in table 1 below

Table 1 Local Weight Formula

Formula	Name	Abbr.
$\frac{tf_{ik}}{\max tf_{ik}}$	Normalized Freq	FREQ
$1+\log tf_{ik}$	Normalized Log	LOGN
$k+\frac{(1-k)tf_{ik}}{\max tf_{ik}}$	Augmented Normalized Term Frequency	ANTF

First approach, one can calculate the term frequency for the word as the ratio of number of times the word occurs in the document to the total number of words in the document. In second approach, logarithm of term frequency is used. Logarithms are a way to de-emphasize the effect of frequency. Logarithms are used to adjust within-document frequency because a term that appears n times in a document is not necessarily n times as important as a term that appears once in that document. Logarithms formulas decrease the effects of large differences in term frequencies. The normalized log (LOGN) given respectively by

$$\begin{aligned} L_{ik} &= 1+\log tf_{ik}\,; \quad if \;\; tf_{ik} > 0 \\ &= \quad 0\;; \qquad if \;\; tf_{ik} = 0 \end{aligned}$$

In third approach, augmented normalized term Frequency give credit to any word that appears and then give some additional credit to words that appear frequently. The formula gives a value of K = 0:5 for appearing in the document plus additional weight that depends on the frequency. This formula was proposed by Croft and parameterized by a value equal to K. Croft suggested that K must be set to something low (0.3) for large documents and to higher values (0.5) for shorter documents. With this formula, the output value varies only between 0.5 and 1 for terms that appear in the document. By restricting the term frequency to a maximum value of 1.0, this technique compensates the problem of the presence of higher term frequencies for normalization.

4.2 Global Weight

Global weight [13] [14] [15] evaluation tries to give a discrimination value to each term. The basic idea is that the term with less frequency appears in the whole collection, it is considered as more discriminating. A commonly used global weight is the inverse document frequency measure i.e. IDF. The table 2 shows the approaches used for global weight.

There are two variations, IDF and IDFP, given respectively by IDF is the logarithm of the inverse of the probability that term i appears in a random document. IDFP is the logarithm of the inverse of the odds that term i appears in a random document. IDF and IDFP are similar in way that they both award high weight for terms appearing in few documents in the collection and low weight for terms appearing in many documents in the collection; however, they differ because IDFP actually awards negative weight for terms appearing in more than half of the documents in the collection, and the lowest weight IDF gives is one.

Table 2 Global Weight Formula

Formula	Name	Abbr.
$log\frac{N}{n_i}$	Inverse Document Frequency	IDF
$log\frac{N-n_i}{N}$	Probabilistic Inverse Document Frequency	IDFP

Where N is the number of documents in the collection and n_i is the number of documents in which term i appears.

4.3 Normalization

Information retrieval approach will have to deal with document with varying length. The third component of the weight evaluation scheme is the normalization factor which is used to correct discrepancies in document lengths. It removes the advantage that long document have over short documents. It is useful to normalize the document vectors so that documents are retrieved independent of their lengths. The most commonly used normalization form in vector space model is cosine normalization. There are two reasons to use normalization. One, same term appears in documents repeatedly. As a result, the term frequency may be large for long documents. Second, long documents also have different numerous terms. This increases the number of matches between a query and a long document. This increases the chances of retrieval of long documents over shorter documents.

5. RESULT

Vector space model and N-layer vector space model are implemented in java. For comparing these two models, the web dataset MathWebPageCorpus [16] from national university of Singapore is used. Using Different weight evaluation scheme, the similarity between the documents and each query is computed in the test collection and returned a list of documents ranked in order of their similarity scores. Different weight evaluation schemes are considered using different local weight and global weight. These weight evaluation schemes are given in table 3.

Same set of queries are fired for all the above method. The precision and recall rate is used to evaluate the performance of vector space model and N-level vector space model. Precision indicate proportion of items retrieved that are relevant and recall indicates proportion of relevant items that are retrieved.

$$Precision = \frac{\text{Number of retrieved relevant document}}{\text{Total number of retrieved document}}$$

$$Recall = \frac{\text{Number of retrieved relevant document}}{\text{Total number of relevant document}}$$

Table 3 different weight evaluation scheme

Method	**Local weight**	**Global Weight**	**Document Normalization**
Cosine Similarity	TF	IDF	cosine Normalization
Normalized term frequency-inverse doc weight (M1)	FREQ	IDF	cosine Normalization
Normalized term frequency probabilistic inverse doc weight (M2)	FREQ	IDFP	cosine Normalization
Augmented term frequency- inverse doc weight (M3)	ANTF	IDF	cosine Normalization
Augmented term frequency- probabilistic inverse doc weight (M4)	ANTF	IDFP	cosine Normalization
log term frequency-inverse doc weight (M5)	LOGN	IDF	cosine Normalization
log term frequency-probabilistic inverse weight (M6)	LOGN	IDFP	cosine Normalization

The result obtained by executing same set queries for above all method. The average precision and average recall is calculated and compared. The average precision and average recall obtained for all seven methods is shown below in table 4.

Table 4: Weight Evaluation Scheme comparison

Method	Average Precision		Average Recall	
	VSM	N-Layer VSM	VSM	N-Layer VSM
Cosine Similarity	0.60	0.66	0.69	0.82
Normalized term frequency-inverse doc weight (M1)	0.59	0.66	0.70	0.82
Normalized term frequency -probabilistic inverse doc weight (M2)	0.59	0.63	0.69	0.80
Augmented term frequency- inverse doc weight (M3)	0.32	0.44	0.43	0.63
Augmented term frequency- probabilistic inverse doc weight (M4)	0.32	0.42	0.43	0.61
log term frequency -inverse doc weight (M5)	0.42	0.48	0.61	0.78
log term frequency -probabilistic inverse weight (M6)	0.43	0.48	0.62	0.77

Fig.1 shows the comparison of average precision of vector space model and N-layer vector space model. The graph clearly shows that average precision of N-layer vector space model is better than vector space model for all the method.

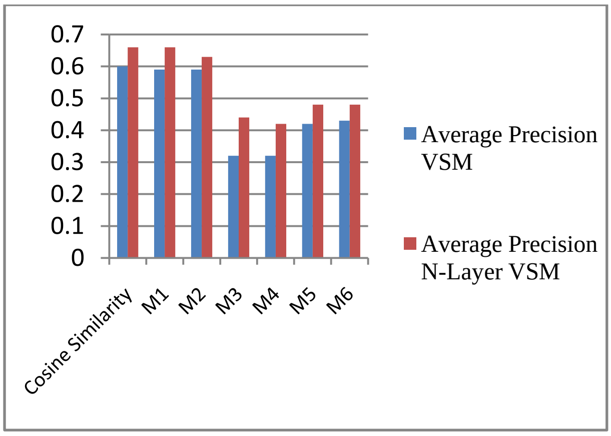

Fig.1 Comparison of average precision of VSM and N-layer VSM

Fig.2 shows the comparison of average recall of vector space model and N-layer vector space model. The graph clearly shows that average recall of N-layer vector space model is better than vector space model for all the method.

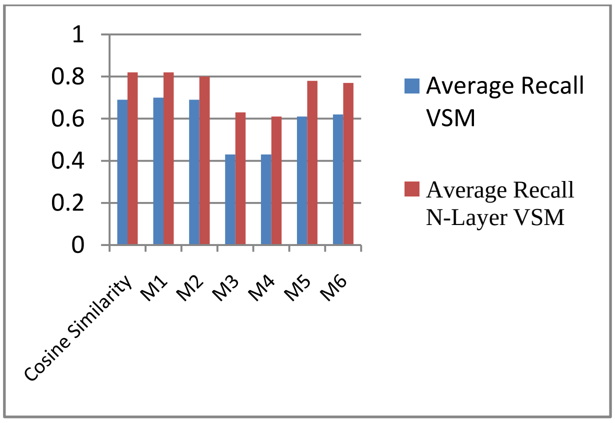

Fig.2. Comparison of average recall of VSM and N-layer VSM

6. CONCLUSION

Term weight evalution is an important aspect of information retrieval systems. Terms are words, phrases, or any other indexing units used to identify the contents of a text. Since different terms have different importance in a text, weight is associated with every term. Proper term weight evalution can greatly improve the performance of the vector space method.

The cosine similarity and six different methods that are formed using local weight and global weight are implemented for vector space model N-layer vector space model. In Cosine Similarity approach, the average precision and average recall of vector space model is 0.59 and 0.69 respectively whereas in N-layer vector space model, the average precision and average recall is 0.66 to 0.82.

In normalized term frequency-inverse doc weight method and normalized term frequency- probabilistic inverse doc weight method, normalized term frequency is used as local weight. In normalized term frequency-inverse doc weight method and normalized term frequency-probabilistic inverse doc weight method, N-layer vector space model gives better result in terms average precision and average recall. The average precision and average recall improved approximately by 10 percent.

In Augmented term frequency-inverse doc weight method and augmented term frequency- probabilistic inverse doc weight method, augmented normal term frequency is used as local weight. In augmented term frequency-inverse doc weight method and augmented term frequency- probabilistic inverse doc weight method, with N-layer vector space model, average precision is improved by approximately by 10 percent whereas average recall is improved approximately by 20 percent.

In log term frequency-inverse doc weight method and Log term frequency-probabilistic inverse weight method, normalized log term frequency is used as local weight. N-layer vector space model does not show significant improvement in average precision. The average precision is improved merely by 4 percent. The average recall is improved approximately by 11 percent.

N-layer vector space model gives better result as compare to vector space model. Cosine similarity and all six weight evaluation methods that formed using different local weights and global weights shows that average precision and average recall in case of N-layer vector space model is always better than vector space model.

7. REFERENCES

[1] Srinath Sriniwas, P.C. Bhatt (2002) "Introduction to Web Information Retrieval: A User Perspective" Resonance June 2002 Resonance, June 2002 Page 27-38

[2] P. Ravikumar, Ashutosh kumar singh (2010) "Web Structure Mining: Exploring Hyperlinks and Algorithms for information Retrieval" American Journal of Applied Science 7(6) 2010 Page 840-845

[3] Anwar A. Alhenshiri " Web Information Retrieval and Search Engine Techniques" Al-Satil Journal Page 55-81

[4] Mehran Sahami, Vibhu Mittal, Shumeet Baluja, Henry Rowley. "The Happy Searcher: Challenges in Web Information Retrieval" Google Inc. 1600 Amphitheatre Parkway, Mountain View, CA 94043

[5] Ricardo Baeza-Yate "Information retrieval in the Web: beyond current search engines" International Journal of Approximate Reasoning 34 (2003) 97–104

[6] Cheng Xiang Zhai, "Statistical Language Models for Information Retrieval A Critical Review" Foundations and Trends in Information Retrieval Vol. 2, No. 3 (2008) 137–213

[7] Joon Ho Lee, "Properties of Extended Boolean models in information Retrieval" Korea research and development center, koera institute of science and technology

[8] http://www.miislita.com/information-retrieval-tutorial/latent-semantic-indexing-fast-track-tutorial.pdf visited on 10/10/2011

[9] Kirk Baker, "Singular Value Decomposition Tutorial"

[10] Norbert Fuhr, "probabilistic model in information retrieval".

[11] Dr. E. Garcia, "A Tutorial on the Robertson-Sparck Jones Probabilistic Model for Information Retrieval"

[12] http://snowball.tartarus.org/algorithms/porter/stemmer.html

[13] G. Salton and C. Buckley. "Term weighting approaches in automatic text retrieval". Information Processing and Managemen 24(5):513{523, 1988.

[14] Christopher D. Manning, Prabhakar Raghavan , Hinrich Schütze, "Introduction to Information Retrieval" Cambridge University Press. 2008.

[15] Ronan Cummins · Colm O'Riordan "Evolving local and global weighting schemes in information retrieval" Inf. Retrieval (2006) 9:311–330

[16] http://wing.comp.nus.edu.sg/downloads/mwc/ visited on 06/12/2012